Guest Lecture for 6.869 Advances in Computer Vision

Activity Recognition

Bolei Zhou
MIT CSAIL
Challenge for Image Recognition

• Variation in appearance.
Challenge for Activity Recognition

- Describing activity at the proper level

  Image recognition?
  No motion needed?

  Skeleton recognition?
  Which activities?
Challenge for Activity Recognition

- Describing activity at the proper level

A chain of events

Making chocolate cookies
Outline

• Video Recognition Datasets
• Video Recognition Models

A little bit about my recent work:
• Temporal Relational Reasoning in Videos
Video Recognition Datasets

• Review on image datasets
Video Recognition Datasets

KTH: 6
UCF: 101
HMDB: 51
ActivityNet: 200
Kinetics: 400
Moments: 339
Video Recognition Datasets

Two video collection methods:
- Collect videos from the web (Youtube, Flickr, etc)
- Crowd-sourcing video collection.
Video Recognition Datasets

• KTH Dataset: recognition of human actions
• 6 classes, 2391 videos

Recognizing Human Actions: A Local SVM Approach. ICPR 2004

https://www.youtube.com/watch?v=Jm69kbCC17s
Video Recognition Datasets

- UCF101 from University of Central Florida
- 101 classes, 9,511 videos in training


https://www.youtube.com/watch?v=hGhuUaxocIE
Video Recognition Datasets

• Kinetics from Google DeepMind
• 400 classes, 239,956 videos in training

https://deepmind.com/research/open-source/open-source-datasets/kinetics/
Video Recognition Datasets

- Moments from MIT
- 1 million 3-second video from 339 generic actions

http://moments.csail.mit.edu/index_test.html
Video Recognition Datasets

• Charades dataset: Hollywood in Homes
• Crowdsourced video dataset

http://allenai.org/plato/charades/
Video Recognition Datasets

- Charades dataset: Hollywood in Homes
- Long chain of actions

[Image of 1st, 2nd, and 3rd generation of datasets]

https://www.youtube.com/watch?v=x9AhZLDkbyc
Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. ECCV’16
Video Recognition Datasets

- Charades dataset: Hollywood in Homes
- Crowd-sourced video dataset

Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. ECCV’16
Video Recognition Datasets

• Charades dataset: Hollywood in Homes
• Demo video

https://www.youtube.com/watch?v=x9AhZLDkbyc
Video Recognition Datasets

• Something-Something dataset: human object interaction
• 174 categories: 100,000 videos

- Holding something
- Turning something upside down
- Turning the camera left while filming something
- Opening something

Poking a stack of something so the stack collapses
Plugging something into something

https://www.twentybn.com/datasets/something-something
Crowd-sourcing Video Collection

https://www.twentybn.com/datasets/something-something
Something-to-Something

Video = Sequence of RGB images

How to represent temporal information?
• Capture the temporal dependency
• Efficiency: 1min 25fps video = 1500 images
Video Recognition Models

• Pre-Deep learning era
  Optic flow, trajectories, bag of words.

• Deep learning era
  Neural Networks
Pre-deep learning Activity Recognition

- Optic Flow: the displacement of pixels
- Gesture lecture by Ce Liu next week on motion estimation
Motion Representations in Activity Recognition

• Optic Flow

https://www.youtube.com/watch?v=JSzUdVBmQP4
Motion Representations in Activity Recognition

• Trajectories: key-point tracking over frames

https://www.youtube.com/watch?v=YN2lDqXz-uc
Motion Representations in Activity Recognition

Improved Dense Trajectory (iDT)

• Global motion compensation (camera motion removal)
• Features from trajectories and HoG
• Bag of trajectories + Fisher Vector + PCA
Deep Learning Models for Activity Recognition

- RGB frame fusion network
- 2-stream network
- 3D convolution network
- Temporal segment network
Single-frame image model

Large-scale Video Classification with Convolutional Neural Networks, CVPR 2014
Performance on the UCF101

<table>
<thead>
<tr>
<th>Method</th>
<th>Spatial ConvNets</th>
<th>Temporal ConvNets</th>
<th>Two-Stream</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td>72.7%</td>
<td>81.0%</td>
<td>87.0%</td>
</tr>
</tbody>
</table>
Multi-frame fusion model
Multi-frame LSTM fusion model

Long-term Recurrent Convolutional Networks for Visual Recognition and Description. CVPR 2015
LSTM: recursive neural networks

- Video Captioning


Sequence to Sequence – Video to Text
2-Stream Network
3D convolutional Networks

Computationally expensive, and a lot of model parameters

If it is RGB frame rather than grey frame, it is actually 4D convolution.
\[ H \times W \times C \times T \]

Learning Spatiotemporal Features with 3D Convolutional Networks, ICCV 2015
3D convolutional Networks

• 3D filters at the first layer.
Summary of Video Recognition Networks

Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. ICCV 2017
Pose Estimation in Videos

https://github.com/ZheC/Realtime_Multi-Person_Pose_Estimation

Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. CVPR’17
Pose Estimation in Videos

Demo Video:
https://www.youtube.com/watch?v=pW6nZXEwIqM&t=77s

https://github.com/ZheC/Realtime_Multi-Person_Pose_Estimation

Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. CVPR’17
Some of my latest work:

Temporal Relational Reasoning in Videos

Bolei Zhou, Alex Andonian, Antonio Torralba
CVPR’18 submission
Temporal Relational Reasoning

• Infer the temporal relation between frames.

Poking a stack of something so it collapses
Temporal Relational Reasoning

- It is the **temporal transformation/relation** that defines the activity, rather than the appearance of objects.

  Poking a stack of something so it collapses
Relational Reasoning for Visual Question Answering

**Original Image:**

**Non-relational question:**
What is the size of the brown sphere?

**Relational question:**
Are there any rubber things that have the same size as the yellow metallic cylinder?
Relational Reasoning for Visual Question Answering

Temporal Relations in Videos

Pretending to put something next to something

2-frame relations

3-frame relations

4-frame relations
Framework of Temporal Relation Networks

Pretending to put something next to something
Something-Something Dataset

• 100 K videos from 174 human-object interaction classes.
  
  Moving something away from something

  Plugging something into something

  Pulling two ends of something so that it gets stretched
Jester Dataset

- 140 K videos from 27 gesture classes.
  - Zooming in with two fingers
  - Thumb down
  - Drumming fingers
# Experimental Results

- On Something-Something dataset

## Model Performance

<table>
<thead>
<tr>
<th>Model</th>
<th>Top1 acc. (%)</th>
<th>Top5 acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>single frame</td>
<td>11.41</td>
<td>33.39</td>
</tr>
<tr>
<td>2-frame TRN</td>
<td>22.23</td>
<td>48.80</td>
</tr>
<tr>
<td>3-frame TRN</td>
<td>26.22</td>
<td>54.15</td>
</tr>
<tr>
<td>4-frame TRN</td>
<td>29.83</td>
<td>58.21</td>
</tr>
<tr>
<td>5-frame TRN</td>
<td>30.39</td>
<td>58.29</td>
</tr>
<tr>
<td>7-frame TRN</td>
<td>31.01</td>
<td>59.24</td>
</tr>
<tr>
<td>MultiScale TRN</td>
<td>33.01</td>
<td>61.27</td>
</tr>
<tr>
<td>MultiScale TRN (10-crop)</td>
<td><strong>34.44</strong></td>
<td><strong>63.20</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Top1 acc. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yana Hasson</td>
<td>25.55</td>
</tr>
<tr>
<td>Harrison.AI</td>
<td>26.38</td>
</tr>
<tr>
<td>I3D by [8]</td>
<td>27.23</td>
</tr>
<tr>
<td>Guillaume Berger</td>
<td>30.48</td>
</tr>
<tr>
<td>Besnet (Top1 on leaderboard)</td>
<td>31.66</td>
</tr>
<tr>
<td>MultiScale TRN</td>
<td><strong>33.60</strong></td>
</tr>
</tbody>
</table>
Experimental Results

- On Jester dataset

<table>
<thead>
<tr>
<th>model</th>
<th>Top1 acc.(%)</th>
<th>Top5 acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>single frame</td>
<td>63.60</td>
<td>92.44</td>
</tr>
<tr>
<td>2-frame TRN</td>
<td>75.65</td>
<td>94.40</td>
</tr>
<tr>
<td>MultiScale TRN</td>
<td>93.70</td>
<td>99.59</td>
</tr>
<tr>
<td>MultiScale TRN (10-crop)</td>
<td><strong>95.31</strong></td>
<td><strong>99.86</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>model</th>
<th>Top1 acc.(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20BN’s Jester System</td>
<td>82.34</td>
</tr>
<tr>
<td>VideoLSTM</td>
<td>85.86</td>
</tr>
<tr>
<td>Guillaume Berger</td>
<td>93.87</td>
</tr>
<tr>
<td>Ford’s Gesture Recognition System</td>
<td>94.11</td>
</tr>
<tr>
<td>Besnet (Top1 on leaderboard)</td>
<td>94.23</td>
</tr>
<tr>
<td>MultiScale TRN</td>
<td><strong>94.78</strong></td>
</tr>
</tbody>
</table>
Experimental Results

• Demo Video:
  http://relation.csail.mit.edu/
Common sense knowledge learned by models
Importance of temporal orders
Activity Forecasting

First Frames

Forecasts

1: Tearing sth just a little bit (0.998)
2: Tearing sth into two pieces (0.001)
3: Pretending to be tearing sth that is not tearable (0.001)

Ground Truth

1: Lifting a surface with sth on it but not enough for it to slide down (0.490)
2: Lifting sth with sth on it (0.423)
3: Tilting sth with sth on it slightly so it doesn’t fall down (0.079)

1: Poking sth so lightly that it doesn’t or almost doesn’t move (0.466)
2: Poking a stack of sth so the stack collapses (0.207)
3: Poking sth so it slightly moves (0.164)

1: Swiping Down (0.881)
2: Swiping Up (0.105)
3: Stop Sign (0.008)
Activity Forecasting

<table>
<thead>
<tr>
<th>Data</th>
<th>Something</th>
<th></th>
<th>Jester</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>baseline</td>
<td>TRN</td>
<td>baseline</td>
<td>TRN</td>
</tr>
<tr>
<td>first 25%</td>
<td>9.08</td>
<td>11.14</td>
<td>27.25</td>
<td>34.23</td>
</tr>
<tr>
<td>first 50%</td>
<td>10.10</td>
<td>19.10</td>
<td>41.43</td>
<td>78.42</td>
</tr>
<tr>
<td>full</td>
<td>11.41</td>
<td>33.01</td>
<td>63.60</td>
<td>93.70</td>
</tr>
</tbody>
</table>
Future Directions in Activity Recognition

How to better model temporal relation?

How to make model more efficient?
- Remove the dependency on optic flow.
- Sampling of discrete frames

Non-local Neural Networks

Xiaolong Wang$^{1,2*}$, Ross Girshick$^2$
$^1$Carnegie Mellon University
$^2$Facebook AI Research

Abstract

Convolutions and recurrent operations are building blocks for processing one local neighborhood at a time. In this paper, we present non-local operations as a generic building block for capturing long-range dependencies, inspired by the classical non-local means method.

Future Directions in Activity Recognition

**Activity forecasting**
What’s next?
kiss, hug, highfive

**Understanding long videos**
Such as movie and TV shows?